A Hierarchical Attention based Model for Off-topic Spontaneous Spoken Response Detection

Andrey Malinin, Kate Knill and Mark Gales {am969,kate.knill,mjfg}@eng.cam.ac.uk

UNIVERSITY OF CAMBRIDGE

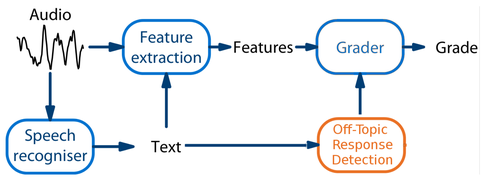
ALTA Institute / Department of Engineering, University of Cambridge

1. Introduction

Need to assess relevance of responses to prompts/topics
- Increases validity of automatic assessment

Solution: Construct relevance assessment system →
- Reject and pass off-topic responses to human graders

Implementation:
- As part of the Grader
- As separate filter stage



Two scenarios - given a set of prompt-response pairs:
- Generative Model of Responses
- Discriminative Model of Relevance

Key Concept: Directly compute $\mathtt{P}(\mathtt{rel}|\boldsymbol{w}^r, \boldsymbol{w}^p)$

2. Attention based Topic Model (ATM)

- Compute a Bi-Directional LSTM prompt embedding:

$$w_1^p \cdots w_L^p \rightarrow \tilde{\boldsymbol{h}}^p$$

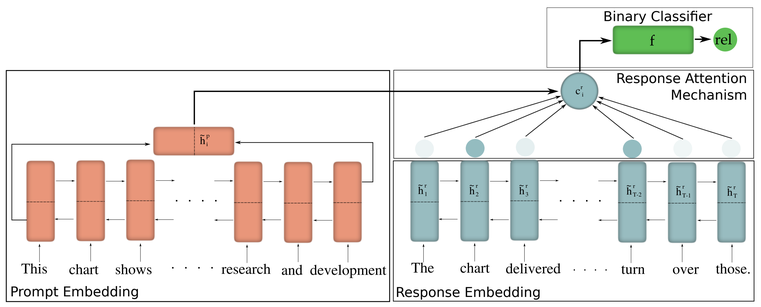
- Encode response as series of BiLSTM response embeddings:

$$w_1^r \cdots w_T^r \rightarrow \tilde{\boldsymbol{h}}_{1:T}^r$$

- Compute prompt-conditional response embedding $\boldsymbol{c}^r$ via Attention
  - Response Attention weights $\alpha_t^r = f(\tilde{\boldsymbol{h}}_t^r, \tilde{\boldsymbol{h}}^p) \rightarrow \sum_{t=1}^T \alpha_t^r = 1$
  - Computes $\boldsymbol{c}^r = \sum_{t=1}^T \alpha_t^r \tilde{\boldsymbol{h}}_t^r$
- Use extracted features $\boldsymbol{c}^r$ to assess relevance:

$$\mathtt{P}(\mathtt{rel}|\boldsymbol{w}^r, \boldsymbol{w}^p) = f(\boldsymbol{c}^r)$$

- Models will require both positive and negative examples for training
- Model can handle arbitrary prompts and responses.

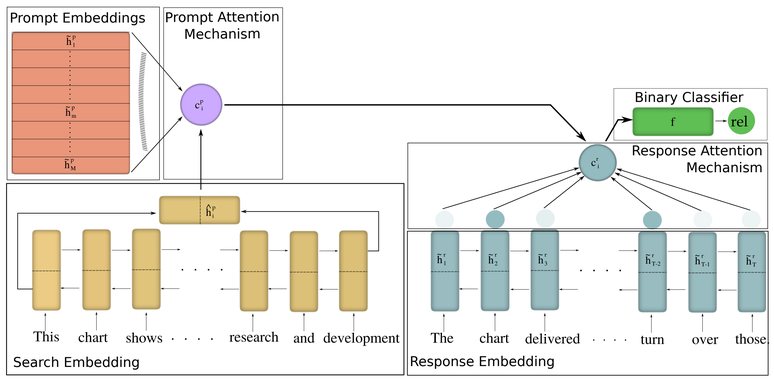


3. Hierarchical Attention based topic model (HATM)

Goal → Improve performance on unseen prompts

Solution → explicitly use train prompts to define topic space
- Express new prompts as interpolation of training prompts
- Prompts projected into topic simplex

Implementation
- Compute a BiLSTM prompt 'search embedding' $w_1^p \cdots w_L^p \rightarrow \hat{\boldsymbol{h}}^p$
- Use $\hat{\boldsymbol{h}}^p$ to attend over training prompts:
  - Prompt Attention weights $\alpha_i^p = f(\tilde{\boldsymbol{h}}_i^p, \hat{\boldsymbol{h}}^p) \rightarrow \sum_{i=1}^M \alpha_i^p = 1$
- Prompt embedding $\boldsymbol{c}^P = \sum_{i=1}^M \alpha_i \tilde{\boldsymbol{h}}_i^p$
- Response Attention weights $\alpha_t^r = f(\tilde{\boldsymbol{h}}_t^r, \boldsymbol{c}^p)$



Prompt attention Entropy is a measure of predictive uncertainty
- Can use to reject to human assessors
- Can use for Active Learning

4. Data

Data	#Topics	#Resp.	#Words	#Resp./ Topic	Avg.Resp. Length
TRN	379	293.0K	13.4M	773.2	45.8
EVAL	219	4077	186.0K	18.6	45.6

- Uniform over CEFR grade levels
- Zipfian over prompts
- ASR %WER per CEFR grade level on EVAL:

A1	A2	B1	B2	C	All
60.3	54.0	44.9	41.8	41.4	45.7

- Data only contains valid responses → generate artificially
  - Randomly match responses to other prompts →
  - Response matched to Prompt→ Positive Response (Example)
  - Response mismatched to Prompt → Negative Response (Example)

5. Experimental Results

Data only contains 'Seen' prompts/responses → unrealistic scenario
- Candidates may give responses relating to 'Unseen' prompts →
- Hold out subset of prompt-response pairs and use as 'Unseen' data

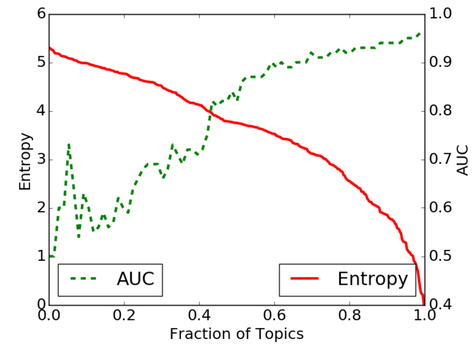
Four Scenarios → different aspects of generalization
- Prompts can be Seen or Unseen
- Negative responses can relate to Seen to Unseen prompts
- Measure performance using ROC AUC

Prompts	System	Neg. Resp. relating to Seen Prompts	Neg. Resp. relating to Unseen Prompts
Seen	ATM	0.949	0.938
	HATM	0.944	0.933
Unseen	ATM	0.855	0.751
	HATM	0.856	0.760

- Model not sensitive to nature of responses given Seen prompts
- For Unseen prompts model performs worse

6. Predictive Uncertainty Evaluation

Scenario - All prompts/topics are seen in training.
- Start with empty EVAL set
- Add pos./neg. examples in order of decreasing prompt entropy.
- Evaluate AUC on each subset of prompts/responses



- Model finds it harder to assess high entropy topics
- AUC increases as fraction of low entropy topics is added to test set

7. Conclusions

- Assess relevance with high AUC between arbitrary prompt-response pairs
- Assess relevance to previously unseen prompts
- HATM can uses prompt attention entropy as uncertainty measure